SI Materials and Methods

Koala samples and datasets

Table 1 lists the koalas sampled. It indicates for each set of koalas the sequencing platform used to generate datasets, the source of the data (generated for the current study or mined from existing data), and also summarizes the results obtained by analysing new datasets or re-analysing previously generated datasets. They are here detailed as follows:

Short read Illumina sequences

Pacific Chocolate (a wild-born New South Wales koala) and Birke (a wild-born Queensland koala) had been sequenced with 100x Illumina short-read coverage (1). Mirali (PCI-SN265) was a zoo koala (northern Australian lineage) from the Vienna Zoo, Tierpark Schönbrunn. It had been Illumina sequenced after KoRV enriched hybridization capture. The dataset is described in reference (2). Archived museum samples of 6 koalas collected in Queensland between 1870 and 1938 had been Illumina sequenced after KoRV enriched hybridization capture, as described in reference (2). All of these koala datasets were examined for the presence or absence of the recKoRV recombination breakpoints shown in **Fig 1**, with results by koala shown in **SI Appendix Fig S1** and **Table S1**.

PacBio long read sequencing

Two koalas were sequenced using the PacBio platform: Bilbo (a wild koala from Upper Brookfield, Queensland) and Bilyarra (from the Tierpark Schönbrunn, Vienna) (**Table 1**). Bilbo is the koala for which the koala reference genome has recently been described (1) Briefly, the long-read genome assembly used in this work is version: phaCin_unsw_v4.1 deposited in DDBJ/ENA/GenBank under the accession GCA_002099425.1 with the genome assembly project registered under BioProject PRJNA359763. High molecular weight (HMW) DNA was extracted from female koala spleen (Australian Museum registration M.47724) using Genomic-Tip 100/G columns (Qiagen) and DNA Buffer set (Qiagen). 15 SMRTbell libraries were prepared and sequenced on the Pacific Biosciences RS II platform (Pacific Biosciences) with a total of 272 SMRT Cells sequenced to give an estimated overall coverage of 57.3X based on a genome size of 3.5Gbp. After filtering low-quality and duplicate reads, approximately 57.3-fold read coverage was used for assembly. Primary contigs (homozygous regions) made up 3.19 GB of the assembled genome, comprising 1,906 contigs, with an N50 of 11.6 Mb and sizes ranging up to 40.6 Mb. Assembly with Falcon (v.0.3.0) including the 5225 alternate contigs of heterozygous regions yielded a 3.42 Gb assembly with an N50 of 48.8 kb. Approximately 30-fold coverage of Illumina short reads were used to polish the assembly with Pilon (1).

Bilyarra (Pci-SN241) of the Vienna Zoo (Tierpark Schönbrunn) died in 2014 and spleen tissue was used to extract DNA for the current study using the QIAamp DNA Minikit (Qiagen). The integration sites and KoRV and recKoRV sequences were enriched prior to PacBio sequencing as described below using Inverse PCR and PacBio sequencing. The data generated from these koalas were of specific relevance to defining full length recKoRV1 in **Fig 1** and integration sites and LTRs described in **SI Appendix, Fig S1 and S2** which could not be accomplished for individual loci using Illumina data.

PCR based screening of recKoRV1 from 166 koala DNA samples

The 166 wild koala DNA samples used to define the distribution of the recKoRV1 3' breakpoint across Australian koala populations were collected by Joanne Meers and Paul Young and their associates (3). Screening of these samples was performed on DNA extracted using the Blood & Tissue DNA Extraction Kit (Qiagen) or from DNA provided by collaborators and tested for integrity using a control primer pair specific for the koala *actin* gene. The study was conducted in accordance with the following permits and approvals: the University of Queensland Animal Ethics Committee (approval numbers SVS/492/12/ARC/WWW; SVS/488/09/ARC/WWW & MICRO/PARA/612/08/ARC); the South Australia Department of Environment, Water and Natural Resources Permit to Undertake Scientific Research (permit numbers A25844; U25790); the South Australia Department of Environment, Water and Natural Resources Wildlife Ethics Committee (approval numbers 12/2010; 51-2009-M1); the South Australia Department of Environment, Water and Natural Resources Export Protected Animals Permit (permit number E20833); Queensland Environmental Protection Agency Wildlife Movement permits (numbers WIWM08219010; WIWM09103211; WIWM09434211; WIWM12645213; WIWM06555009; WIWM06798010), NSW National Parks and Wildlife Service Export Licence, number IE106347.

PCR screening of recKoRV1 from the 166 koalas (**Fig 2**) involved two primer sets (Ya_recKorRV1-F: 5'-GCT GCT TGA TTT GGA TGT GA-3'; Ya_recKoRV1-R: 5'- GAG GAG TAG CAG GGG ACC AG-3'; recKoRV-F1: 5'-TGT GAA TAT CCC TGG CAG CCG CG-3'; KoR27-R: 5'-GAG TAA CAG AAG GAG GAG TAG CAG-3'). Sequences were trimmed and visualized using the alignment and assembly program Vector NTI advance 11 (Invitrogen). It should be noted that using this PCR strategy, recKoRV1 and recKoRV2 cannot be readily distinguished (4). However, recKoRV2 was generally rare in both the koala reference genome and in Pci-SN241, and the two recombinants likely are very closely related so this would not change the interpretation of the results. The resulting data is presented in **Fig. 2**.

Inverse PCR and PacBio sequencing of Bilyarra to determine KoRV and recKoRV sequences

The protocol employed has four steps: (1) the DNA must be randomly fragmented, (2) the fragmented DNA is end repaired and ligated into closed circles, (3) LTR primed long fragment PCR is performed on the closed circles, (4) The resulting KoRV/recKoRV integration enriched PCR products are PacBio sequenced.

Step 1, fragmentation: DNA extracts from Pci-SN241 were quantified on the Qubit Fluorometer (high sensitivity chemistry) as well as the 2200 TapeStation (Agilent Technologies) using the Genomic DNA (gDNA) Screen Tape, showing DNA size to be primarily distributed around 50-60 kb. To convert the DNA to a size suitable for this study (average of 3 to 4 kb), DNA extracts were diluted to 50 ng/µL in a final volume of 200 µL. The extracts were fragmented on the Covaris M220 in a miniTUBE Blue using the following settings (in parentheses): intensity (0.1-0.5), duty cycle (20%), cycles per burst (1000), total treatment time (600 seconds), temperature (20°C). DNA fragmentation profiles of the sheared DNA were further assessed on the TapeStation using a gDNA Screen Tape. After confirming that the size of the sheared DNA was between 2-7 kb, 42.5 µL of the sheared DNA was used in blunt end reactions run in triplicate using the Fast DNA End Repair Kit (Thermo Scientific). The products were purified using the QIAquick PCR Purification Kit (Qiagen, Hilden, Germany) and were again quantified on the TapeStation. Because the results were similar across the triplicate runs, they were pooled together.

Step 2, Circularization: To find the optimal ligation conditions for subsequent inverse PCR, ligations were performed using a series of varying (total) input blunt ended DNA. The following amounts were tested: 5ng, 10ng, 15ng, 25ng, 30ng, 40ng, 50ng, 75ng and 100ng, each in a 50µL ligation reaction. Ligation reactions used a T4 DNA Ligase kit (5 U/µL) (Thermo Scientific) with 5 µL of T4 DNA Ligase Buffer (10X), 5 µL of 50% PEG 4000 solution, 2.5 µL of T4 DNA Ligase and an amount of blunt ended DNA (as indicated by the series), and molecular biology grade water to a 50µL total volume. Ligation was performed in a thermal cycler at 16˚C for 16 hours followed by enzyme inactivation at 70˚C for 5 min. Given the minuscule starting DNA amounts, all ligations were performed in triplicate. The ligation products were measured on the TapeStation, showing a 2kb size shift towards higher molecular weight compared to the blunt ended DNA, a sign of transformation of DNA structure from linear to circular. Results for each of the ligation reactions using the same amount of DNA were similar in profile as measured on the TapeStation, so the products of each three were pooled together. A partial ligation (circularization) gradient (25ng, 30ng, 40ng, 50ng, 75ng) was rerun to test for reproducibility, with comparable results.

Step 3, Long Inverse PCR: KoRV proviral genomes were downloaded from GenBank (accessions: KF786280, KF786281, KF786282, KF786283, KF786284, KF786285, KF786286, AB721500, KC779547) and aligned using the MAFFT plugin in Geneious version 7.1.7 using default settings (5). For inverse PCR, one set of primers was designed using Primer3Plus software (6) targeting a conserved region in the middle of the KoRV LTR (iPCR_LTR_F; TGCATCCGGAGTTGTGTTCG; iPCR_LTR_R: AAAAGCGCGGGTACAGAAGC).

To avoid loss of circularized DNA during purification, circularization products were directly taken as template without purification for inverse PCR (7). A total of 10ng (as quantified by the Tapestation) from each product in the circularization gradient was taken as template in a separate inverse PCR. The template was amplified using the MyFi Mix (Bioline GmbH, Luckenwalde, Germany) with thermal cycling conditions of an initial denaturation step at 95°C for 1 min 30 sec; followed by 35 cycles at 95°C for 20 sec, 59°C for 20 sec and 72°C for 5 min; final extension of 2 min at 72°C. The products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) and concentration and DNA profiles measured on the TapeStation. The optimal circularization product in each gradient was chosen by considering (i) the DNA amount per microliter of inverse PCR product in the 2-7 kb range, (ii) the average length distribution between 600 bp – 7 Kb range, and (iii) the percentage of DNA within the 2-7 Kb range. Based on these criteria, 40ng of input DNA (conc. 0.8ng/µL in circularization) was used.

To test whether increasing template amount (circularization product) for inverse PCR would affect the length distribution of the PCR product, a series of template amounts (2 ng, 6 ng, 10 ng, 14 ng, 16.8 ng which is the amount of DNA in a maximum input volume of 21 µL) from the two optimal circularization products were used for inverse PCR using same kit and protocols described above. All products were purified using the QIAquick PCR purification kit (Qiagen, Hilden, Germany) and were measured on the TapeStation. The optimal inverse PCR product amount in the series was chosen based on the three criteria above and on the overall distribution of the fragment size peaks determined by the TapeStation measurement. Following this analysis, the inverse PCR with optimal template amount (6 ng) was repeated

twice. These three optimal products were then pooled for PacBio sequencing to minimize clonal PCR bias.

Step 4, PacBio sequencing: PCR products were submitted for PacBio library construction and sequencing to the Max Delbrück Center, Berlin. PCR products were purified using AMPure XP beads (Beckman Coulter), first at a concentration of 0.4X followed by a subsequent purification of the supernatant at 0.6X. The resulting four samples were prepared as sequencing libraries using the PacBio (Pacific Biosciences, Menlo Park, CA) 5kb template prep protocol and the SMRTbell™ Template Prep Kit 1.0 following the manufacturer's recommended protocols. Library concentration and fragment length were verified using the Qubit 2.0 fluorometer (Life Technologies) and the 2100 Agilent Bioanalyzer, using the 12000 DNA chemistry (Agilent Technologies). The estimated average lengths for the short and large insert libraries were 1600 bp and 3500 bp, respectively. Sequencing on the PacBio RSII platform used the MagBead Standard protocol, C4 chemistry and P6 polymerase on a single v3 Single-Molecule Real-Time (SMRT) cell with 1x180 min movie for each library (a total of 4 libraries). The reads from the insert sequence were processed within the SMRTPortal browser (minimum full pass = 1; and a minimum predicted accuracy of 90).

Amplification from integration site to integration site for 11 loci identified four of the loci as being different recKoRVs, i.e., they had recombination breakpoints that differed from recKoRV1 (**Table S3**). Most sequences that turned out to be recKoRVs other than recKoRV1 mapped relatively poorly initially to recKoRV1 whereas those that were confirmed mapped well. Sequencing of the 5' breakpoint was particularly difficult due to large numbers of homopolymer stretches, and several products could not be sequenced. Three loci could not be amplified, likely due to the low complexity sequences flanking the integrations and the difficulty of amplifying a 6.4 kb product in the presence of the empty site on the opposing chromosome. After of mapping and Sanger sequencing, of 14 integrants putatively identified as recKoRV1, ten were confirmed to be recKoRV1..

Bioinformatics analyses of of PacBio sequences to identify integration sites and determine whether the integrants were KoRV or recKoRV proviruses

To isolate the host genomic sequences flanking integration sites for KoRV and recKoRV, the KoRV containing reads were aligned using blastn to the KoRV-A or -B reference sequences (AB721500.1; KC779547). Regions homologous to the reference sequences were removed. The isolated host genomic sequences flanking integrations sites were clustered using Tribe-MCL (I=1.4) (8), a Markov cluster based approach, processing distance based information of a blastn matrix for all KoRV containing reads (8). The recKoRV1 containing reads were aligned using blastn to the KoRV-A and -B reference sequences, as well as to PhER, all known recKoRV breakpoints and the consensus sequence of recKoRV1. Regions homologous to any of the reference sequences were removed. The isolated flanking regions were clustered using Tribe-MCL (I=4). A consensus sequence for every cluster was created by constructing a multiple sequence alignment using MAFFT (v7.305b) (9) and computing a consensus sequence using the Perl module BioPerl::SimpleAlign (30 % identity, gap removal) (10).

Raw Bilyarra circular consensus sequences (ccs), KoRV and recKoRV1 insertion site flanking sequences (a consensus of all sequences) were mapped to the assembled genome of koala Bilbo using the Burrows-Wheeler alignment (BWA BWA-SW default) for long sequences. Regions of interest were determined using Bedtools (11, 12). We examined regions covered by at least 30 ccs in Bilyarra, and the regions of interest were manually

annotated. Each KoRV or recKoRV1 integration site determined using the described bioinformatics approaches was then confirmed by PCR including primers based on the regions flanking integration sites, using Sanger sequencing to determine whether the elements and structures identified were consistent with the bioinformatics analysis of the inverse PCR products (**SI Appendix, Fig S3, Table S3**). The results of this analysis are the basis of **SI Appendix, Table S1**, **S2**. Both Bilbo and Bilyarra had KoRV and recKoRV integrations with the 5' and 3' LTRs belonging to different LTR groups suggesting that gene conversion or recombination, both observed in other proviruses, had occurred, precluding accurate dating of individual integrations.

Network analysis of PacBio generated KoRV and recKoRV LTRs
LTRs from all insertion sites of Bilbo and Bilyarra were aligned with sequences of LTRs from KoRV integrations in ten different koalas examined in (13), and with KoRV-A (AB721500.1) and KoRV-B (KC779547.1). The iPCR primer gaps were removed in all sequences. Multiple sequence alignment was performed using MAFFT L-INS-i (14). The alignment was cropped to the most conserved regions (>89% identity) on both ends, realigned and manually curated. A haplotype network was constructed using the R (15) package Pegas (16) with the distance model "indelblock", performing an iterative refinement for the smallest sum of distances. The results of the analysis are shown in **SI Appendix, Fig S2**.

Bioinformatics analysis of koala Illumina sequence data
Hybridization capture of KoRV is described in (2). The recKoRV1 breakpoint sequences in Pacific Chocolate and Birke were initially detected in transcriptome sequences (17). Subsequently, Illumina 100 bp genomic sequence libraries from both of these koalas (1) were screened. First, a subset of reads enriched in KoRV sequences was produced using the fastmap mode of bwa to align reads to a reference KoRV genome sequence. Second, using blastn the KoRV-enriched set of reads was filtered to remove reads with full-length alignments to KoRV, leaving reads of potentially chimeric sequences. Finally, blastn was used to query these potential chimeric reads to a PhER sequence.

Next generation sequence data from archival samples, obtained from (2), were filtered using cutadapt v1.8.1 for adaptor sequence, low quality reads, and fragments shorter than 30 bp (18). Sequence data that passed the quality filters were aligned to breakpoints identified in recKoRV1 using BWA version 0.7.15-r1140 and the mem algorithm with default settings (19). Aligned data were further processed using samtools for clonal read removal. Identified breakpoints were confirmed visually using Geneious 7.1(5). The results of breakpoint characterization are shown in **Fig 1** and **SI Appendix**, **Fig S1** and **Table S1**.

Koala transcriptome analysis
Expression of PhER was tested by searching for PhER sequences in the transcriptomes reported by Hobbs et al. (2014) (17). An 8 kb PhER sequence (Hobbs et al., 2017) (4) was used as a query in blastn searches of Pacific Chocolate and Birke transcriptome databases. In both cases, the searches revealed PhER transcripts distinct from those forming part of recKoRV, and which notably included those parts of PhER that are not incorporated into recKoRV.

Ackowledgments
The authors wish to thank the following researchers who supplied koala DNA, blood or tissue samples to the KoRV research group at the University of Queensland (UQ): Bill Ellis, Greg

Simmons and Kirsten Lee (UQ), Jon Hanger and Jo Loader (Endeavour Veterinary Ecology), Sam Gilchrist (Wild Life, Hamilton Island), Michael Pyne (Currumbin Wildlife Sanctuary Hospital), Steve Phillips (Biolink Ecological Consultants), Rodney Starr (Port Stephens Veterinary Services), Mick Murphy (National Parks & Wildlife Service, NSW), Lynley Johnson and Wayne Boardman (ZoosSA, Cleland Wildlife Park), Greg Johnson (Kangaroo Island Veterinary Clinic), Robyn Molsher (South Australia Department of Environment, Water & Natural Resources), Amber Gillett (Australia Zoo Wildlife hospital), Cheyne Flanagan (Port Macquarie Koala Hospital) and Kath Handasyde (University of Melbourne).

SI References

1. Johnson RN, Chen Z, Etherington GJ, Ho, SYW, Nash WJ, *et al.* (2018) Adaptation and conservation insights from the koala genome. *Nature Genetics* In Press.
2. Tsangaras K, *et al.* (2014) Hybridization Capture Reveals Evolution and Conservation across the Entire Koala Retrovirus Genome. *Plos One* 9(4).
3. Simmons GS, *et al.* (2012) Prevalence of koala retrovirus in geographically diverse populations in Australia. *Aust Vet J* 90(10):404-409.
4. Hobbs M, *et al.* (2017) Long-read genome sequence assembly provides insight into ongoing retroviral invasion of the koala germline. *Sci Rep* 7(1):15838.
5. Kearse M, *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647-1649.
6. Untergasser A, *et al.* (2012) Primer3--new capabilities and interfaces. *Nucleic Acids Res* 40(15):e115.
7. Sun W, *et al.* (2013) Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing. *EMBO J* 32(14):2029-2038.
8. Enright AJ, Van Dongen S, & Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30(7):1575-1584.
9. Katoh K & Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30(4):772-780.
10. Stajich JE, *et al.* (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res* 12(10):1611-1618.
11. Li H & Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26(5):589-595.
12. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6):841-842.
13. Ishida Y, Zhao K, Greenwood AD, & Roca AL (2015) Proliferation of Endogenous Retroviruses in the Early Stages of a Host Germ Line Invasion. *Molecular Biology and Evolution* 32(1):109-120.
14. Katoh K & Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.
15. Team RDC (2011) R: A Language and Environment for Statistical Computing. (the R Foundation for Statistical Computing, Vienna, Austria).
16. Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26(3):419-420.

17. Hobbs M, *et al.* (2014) A transcriptome resource for the koala (Phascolarctos cinereus): insights into koala retrovirus transcription and sequence diversity. *BMC Genomics* 15:786.
18. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. . *EMBnet* 17(1):10-12.
19. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.

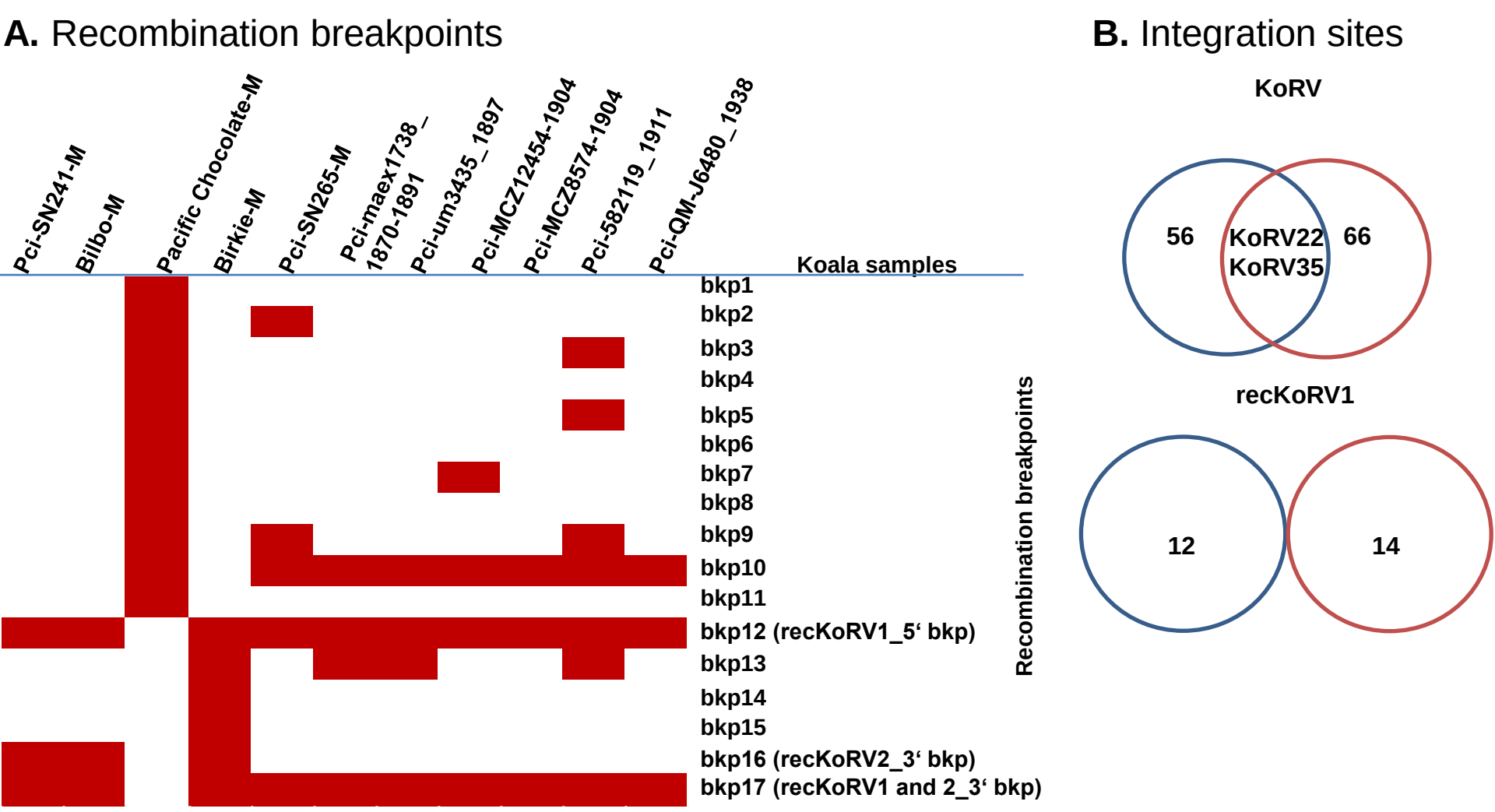

Fig S1. Occurrence of recombination breakpoints and integration sites across koalas. Panel A represents as a heatmap presence (red) or absence (white) data from **SI Appendix, Table S1**, representing identification of recKoRV1 breakpoints (bkps) among koalas: in the inverse PCR PacBio sequenced data Bilyarra, the koala reference genome Bilbo (KGC), Illumina sequenced koala genomes (Pacific Chocolate, PC, and Birke) and hybridization capture data from (22). All koalas are from Queensland except for Pacific Chocolate who derives from New South Wales. Zoo koalas in the study are primarily derived from Queensland populations. Modern samples are labelled with "M" while the date of collection is indicated for museum samples. Recombination breakpoints (bkp) 1-17 are designated on the vertical axis, with those present in recKoRV1-3 specially indicated. In panel B, Venn diagrams indicate the degree of overlap of KoRV and recKoRV1 integrations between Bilbo (blue, the reference genome from a Queensland koala) and Bilyarra (red; from the Vienna Zoo); only 2 of 120 KoRV integration sites were shared between the two koalas, and none of the 26 recKoRV1 integration sites were shared. The full sequences and reference genome locations for all shared and unique integrations are in **SI Appendix, Table S1**.

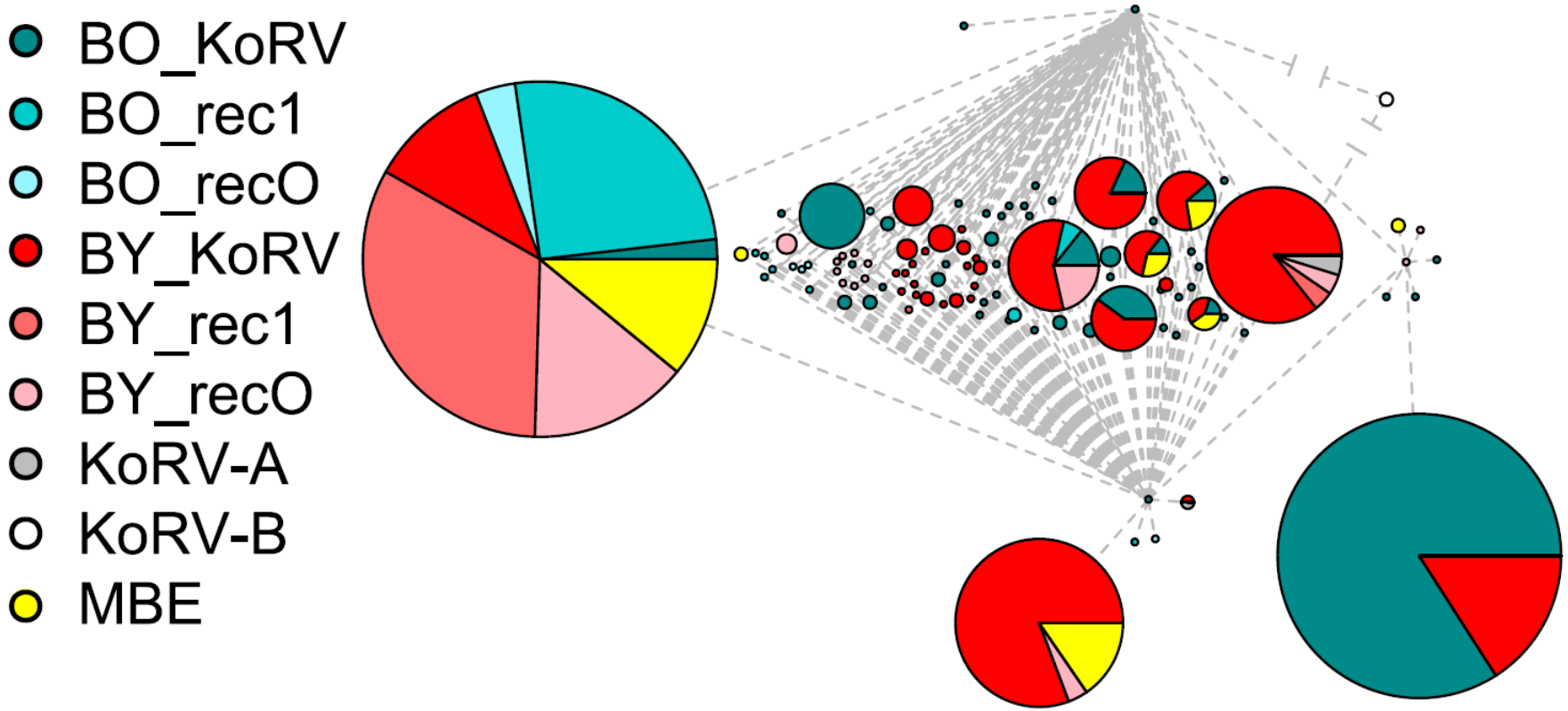

Fig S2: Relationships among KoRV LTR sequences. A minimum spanning network shows the relationships among the KoRV, recKoRV1 (rec1) and other recKoRV (recO) LTR sequences identified in koalas Bilyarra (BY), Bilbo (BO) and koalas (MBE) examined in reference (20) Each pie chart represents a distinct LTR sequence, with circle sizes proportional to the frequency of occurrence of each sequence. Alternative relationships among the network are shown as grey lines. The poor resolution among LTR groups is due to the low diversity among individual elements often differing by single nucleotide differences.

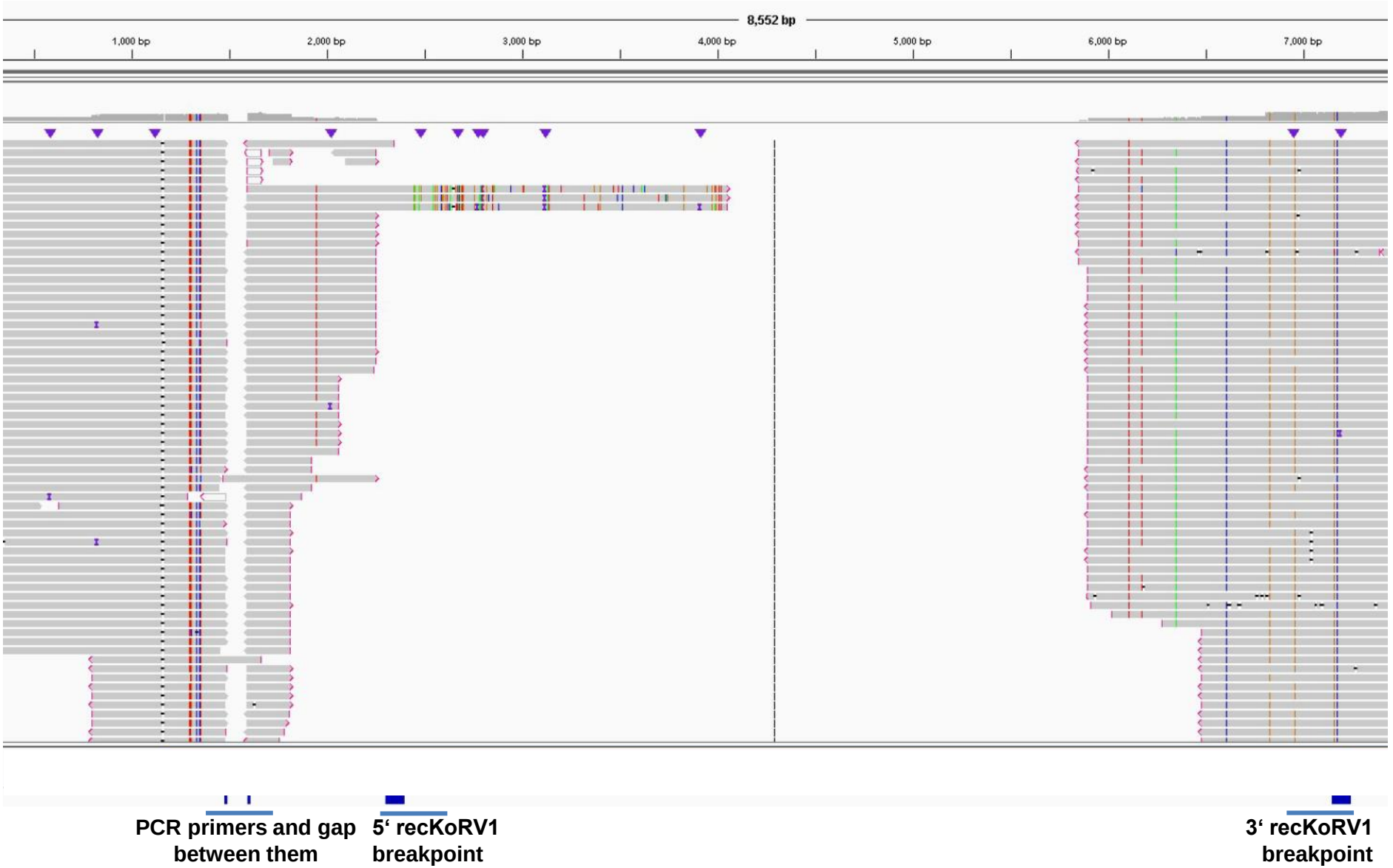

Fig S3. An example mapping of LTR derived inverse PCR PacBio sequences to the recKoRV1 reference. The figure illustrates that fewer reads extended beyond the 5‘ recombination breakpoint than the 3‘ recombination breakpoint of recKoRV1.

Table S1. The recKoRVs identifed in the current study and their distribution in modern and historical koala genomes

Breakpoint name*	Individual	Breakpoint orientation†	KoRV orientation	Breakpoint position in KoRV genome**	Breakpoint position in PhER genome***	Microhomology	Recombination breakpoint sequence‡	Number of reads mapping to breakpoint	Pci-SN265	Pci-QM-J6480	Pci-um3435	Pci-MCZ12454	Pci-MCZ8574	Pci-maex1738	Pci-582119
bkp1	Pacific Chocolate	KoRV->PhER	reverse	6019	4561		gccaggtgagagtcatgggttggtgagggttgttatgcAGATGTGTATAAGAGACAGC AATAATAGGACAAAGGTGTAATGTGATAGATGTCTAATTAAAT	1							
bkp2	Pacific Chocolate	PhER->KoRV	reverse	504/8430	4670		AGAAAAGCAATAGGATTAGACTGATATGATTGGCTTCCAAGATGTATt gaaagatcccaatgttcgggtagtccaccgaccttgagaaaccctcccaggat	6	X						
bkp3	Pacific Chocolate	KoRV->PhER	reverse	3/7929	4667		ggaatgatttctgcctcatgatttctgcctccttcaGTATCTATCATAGCAATCCAAGC TTTAAGCATGTTTAGGTAGGAAGATCACAGTTTTGGATTTT	7							X
bkp4	Pacific Chocolate	PhER->KoRV	reverse	504/8430	5627		GGGGGATTGGCAGTATGGCTCCGGCTCTTATGGCAGTCCTCTTCC TTGGATGATTCtgaaagaccccaaagttcgggtagtccaccgaccttgagaaac	4							
bkp5	Pacific Chocolate	KoRV->PhER	reverse	3/7929	5624		cgggtagtttccatactccacggaatgatttctgcctcatgatttctgcctccttcaATTCCTTTG CCAAAAACTAAAAGGGGGGAGTAGGAACTCAGGG	7							X
bkp6	Pacific Chocolate	KoRV->PhER	forward	504/8430	7310		ggagttgtgttcgcgttgatcctgggagggtttctcaaggtcggtggactacccgaactttgaggtct ttcaCCCCATGGTTAACCACTCTTATTTCTGC	3							
bkp7	Pacific Chocolate	KoRV->PhER	forward	504/8430	29/7585		gttgatcctgggagggtttctcaaggtcggtggactacccgaactttggggtctttcaATTTTAC CAGGTAGTAGAACCCTTTAGGCCATGGAAAACTCCA	4				X			
bkp8	Pacific Chocolate	PhER->KoRV	forward	7448	79/7635	CAAGAC	GGTAGTAGAACCCTTTAGGCCATGGAAAACTCCAGATTCCTTAGcaa gactccataagtaaactagaggattccttaacctccctgtctgaagtagtgctc	19							
bkp9	Pacific Chocolate	KoRV->PhER	forward	3471	174/7731	CCCTC	ctcggctaccgcaaggattcaagaactctcccaccctcAGATCATATCTCTGGAAGC CTCCATTGTGAGAAAGTTTGCAGAGCATTATCATATCTACACAG	1	X						X
bkp10	Pacific Chocolate	KoRV->PhER	forward	1395	309/7869	CCCCTCC	ggagacggacagcctgttcctcctcgcagaaccccgccctatccaacatcccctccCTGT CCTGTATACTCCCTATATAAACCCTAGCTCAA	22	X	X	X	X	X	X	X
bkp11	Pacific Chocolate	PhER->KoRV	forward	5263	348/7908		AAGATCCATCCCCCTCCCTGTCCTGTATACTCCCTATATAAACCCT AGGACAGAacaaaaacagagacagccctgaccgtctgcaagaagatactagagg	5							
bkp12 (recKoRV1, 2 and 3 5' breakpoint)	Birke	KoRV->PhER	forward	1177	3182	AGGAGACT	ttaatccttctattatttctgcagtcaaaaggattgtcttccaggagactgTTGGACCTGGTGA AAGGGTCTAGAAAGAAGAAAGATTTTTTTTATTGGGTC	808	X	X	X	X	X	X	X
bkp13	Birke	KoRV->PhER	forward	504/830	1/7555		tcctgggagggtttctcaaggtcggaggactacccgaacattggggtctttcaATGTAATGC AGAACATTATTTTACCATTTGGGTAGTAGAACCCTTTAG	2			X			X	X
bkp14	Birke	PhER->KoRV	forward	3/7929	73/7629		TAGGTAGTAGAACCCTTTAGGCCATGGAAAACTCCAAATTCCTTAGt gaaggaggcagaaatcatgaggcagaaatcattccgtggagtatggaaactacc	16							
bkp15	Birke	KoRV->PhER	forward	506/8431	66/7622		ctgggagggtttctcaaggtcggaggactacccgaacattggggtctttcattAGCAAGACC CTTATCTGGCTGCTCAGGCTCCAGCCCTCAGATCATATC	19							
bkp16 (recKoRV3 3' breakpoint)	Birke	PhER->KoRV	forward	36/7962	454/8014		TAATAGCTAATAAAGTCCACAATTTAAAATTAGCTCGGGTCGTCAAC TCGCtccgtggagtatggaaactacccggagggcccaaggtttagggacaggtg	60							
bkp17 (recKoRV1 and 2 3'breakpoint)	Birke	PhER->KoRV	forward	7619	446/8006	AGGAGACT	TTTAAAATTAGCTTGGGTCGTCAACTCGCCTATAGGGATGAGAGTG GCCCCaggagactcaaggaaaggttagataagaggcagttagagcaccaaaagaa	660	X	X	X	X	X	X	X

*** Breakpoints 12, 16 and 17 are found in recKoRV1, 2, or 3 (see Figure 1 of the main text)**

† Breakpoint orientations indicate whether KoRV or PhER are 5' or 3' of the recombination breakpoint

‡ KoRV sequences are shown in black and PhER sequences in red capitalized letters

**** 8431 bp; GenBank accesssion AF151794. Two positions given when breakpoint is within LTR.**

***** 8031 bp; positions 10912078-10920108 of scaffold phaCin_unsw_v4.1.fa.scaf00062 (NCBI reference sequence NW_018344013.1) from phaCin_unsw_v4.1 genome assembly. Two positions given when breakpoint is within LTR.**

Koalas in dark grey are modern and those in light grey are historical

Table S2. KoRV and recKoRV1 integration sites identified in Bilyarra and their genomic locations relative to the koala reference genome

Koala	Scaffold number	Retrovirus	Orientation relative to scaffold sequence	Position in scaffold	Target site duplication
Bilyarra	000062F-078-01	KoRV	forward	4230	ND
Bilyarra	000159F-031-01	KoRV	forward	1095	ND
Bilyarra	phaCin_unsw_v4.1.fa.scaf00004	KoRV	forward	23615392	CTTAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00006	KoRV	forward	17890343	ATACTGA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00007	KoRV	forward	1797116	GGCC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00016	KoRV	forward	11903436	CTAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00027	KoRV	forward	2853594	ACCTT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00031	KoRV	forward	6150324	AAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00039	KoRV	forward	17332612	GTTC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00041	KoRV	forward	12435187	GTAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00043	KoRV	forward	10816237	AGAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00055	KoRV	forward	9889936	AGTCCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00058	KoRV	forward	9863928	GATG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00081	KoRV	forward	583341	AGGG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00088	KoRV	forward	1318328	AGAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00088	KoRV	forward	2525660	(ACAC)**
Bilyarra	phaCin_unsw_v4.1.fa.scaf00088	KoRV	forward	4799221	AAAAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00106	KoRV	forward	8525460	TGCCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00127	KoRV	forward	8654899	GTGG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00137	KoRV	forward	5320687	CCAT(g/t)
Bilyarra	phaCin_unsw_v4.1.fa.scaf00137	KoRV	forward	5426576	AAGC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00137	KoRV	forward	7511603	GTAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00150	KoRV	forward	3620463	GGAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00164	KoRV	forward	3258696	GAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00164	KoRV	forward	3402705	(GTGT)**
Bilyarra	phaCin_unsw_v4.1.fa.scaf00164	KoRV	forward	5264524	TAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00228	KoRV	forward	1503766	AATAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00241	KoRV	forward	3345654	CCTG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00273	KoRV	forward	3145264	AGAGG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00322	KoRV	forward	213	GAAGTGA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00354	KoRV	forward	1495511	GAGC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00354	KoRV	forward	1537527	ND
Bilyarra	phaCin_unsw_v4.1.fa.scaf00441	KoRV	forward	282050	ATTC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00001	KoRV	reverse	3902862	GTAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00002	KoRV	reverse	25542095	CCAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00002	KoRV	reverse	32059851	CTAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00008	KoRV	reverse	24417033	AGAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00013	KoRV	reverse	14937508	TAGG/CAAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00031	KoRV	reverse	17598880	AGTACT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00038	KoRV	reverse	7124654	GTATG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00038	KoRV	reverse	12281933	AGGAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00040	KoRV	reverse	13892345	CAAAACC

Bilyarra	phaCin_unsw_v4.1.fa.scaf00048	KoRV	reverse	8799477	GAAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00053	KoRV	reverse	5206411	ATCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00070	KoRV	reverse	7827832	ATACT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00074	KoRV	reverse	2274509	ATAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00082	KoRV	reverse	9179129	ATTAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00094	KoRV	reverse	1633494	GAAA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00098	KoRV	reverse	2205920	GAAA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00107	KoRV	reverse	9030033	AGAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00111	KoRV	reverse	5981086	GTAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00150	KoRV	reverse	2758486	AAAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00159	KoRV	reverse	14458	AGGC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00164	KoRV	reverse	5577461	GCTC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00241	KoRV	reverse	1620865	AGCAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00279	KoRV	reverse	1841947	ATTCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00304	KoRV	reverse	865800	AACT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00310	KoRV	reverse	760620	AGTA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00316	KoRV	reverse	1386589	GTCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00363	KoRV	reverse	1083497	AGAATT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00491	KoRV	reverse	1539	ATTACT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00634	KoRV	reverse	18365	GCTC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00088	KoRV	ND	4803899	ACTCC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00097	KoRV	ND	2135856	ATGA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00166	KoRV	ND	541779	AACAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00218	KoRV	ND	1849618	CAAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00021	recKoRV1	forward	16384538	ACACT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00024	recKoRV1	forward	4934798	CTTA
Bilyarra	phaCin_unsw_v4.1.fa.scaf00083	recKoRV1	forward	797972	ACAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00275	recKoRV1	forward	2910435	AGGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00003	recKoRV1	reverse	3474508	ATAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00037	recKoRV1	reverse	16563123	ATGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00037	recKoRV1	reverse	17582840	GAAG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00069	recKoRV1	reverse	987368	TTGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00079	recKoRV1	reverse	1270278	CCTGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00164	recKoRV1	reverse	6007491	CTTTTT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00002	recKoRV*	forward	23377285	TTAC
Bilyarra	phaCin_unsw_v4.1.fa.scaf00035	recKoRV*	forward	5731424	ATGG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00035	recKoRV*	forward	15145894	CTGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00061	recKoRV*	forward	4737977	(G)CTAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00107	recKoRV*	forward	4792775	AGGGCTG
Bilyarra	phaCin_unsw_v4.1.fa.scaf00139	recKoRV*	forward	4264851	AAGAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00234	recKoRV*	forward	4233389	AAGAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00014	recKoRV*	reverse	15030324	ACCT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00031	recKoRV*	reverse	11457756	CATAAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00060	recKoRV*	reverse	13428694	TGCAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00095	recKoRV*	reverse	9508496	ATAGT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00096	recKoRV*	reverse	1287297	(CTCT)**
Bilyarra	phaCin_unsw_v4.1.fa.scaf00173	recKoRV*	reverse	557197	GTAT
Bilyarra	phaCin_unsw_v4.1.fa.scaf00003	recKoRV*	ND	10330590	TAAC

* KoRV recombinant with PhER, but different breakpoints than recKoRV1. Exact breakpoints could not be determined accurately
** TSD within repetitive, low complexity
region

*** 5' and 3' LTR coul not be determined, since the orientation of the virus relative to the scaffold was not clear

Table S3. Identification of recKoRV1 by mapping and confirmation by integration flanking sequence full proviral genome amplification and Sanger sequencing

Insertion	GAG/PhER			ENV/PhER			recKoRV1
	iPCR 99%	Sanger	bwa	iPCR 99%	Sanger	bwa	
Scaf00003	16	Ø	✓	107	Ø	✓	✓
Scaf00014	0	×	×	1	×	×	×
Scaf00021	4	✓	✓	37	✓	✓	✓
Scaf00024	0	✓	✓	39	✓	✓	✓
Scaf00037_A	6	×	✓	65	✓	✓	✓
Scaf00037_B	22	×	✓	224	✓	✓	✓
Scaf00069	25	×	✓	124	✓	✓	✓
Scaf00079	8	×	✓	46	✓	✓	✓
Scaf00083	15	✓	✓	35	✓	✓	✓
Scaf00096	12	×	×	168	×	✓	×
Scaf00164	14	Ø	✓	136	Ø	✓	✓
Scaf00173	21	×	✓	2	×	×	×
Scaf00234	1	Ø	×	184	Ø	✓	×
Scaf00275	5	×	✓	55	✓	✓	✓

* The symbol Ø indicates loci that could not be amplified by PCR most likely to to the low complexity of the sequences flanking most integrations

** red crosses indicate the inability to sequence through a given breakpoint or failure to map by bwa. This was particularly a problem for the 5' breakpoint which has multiple long homopolymer stretches.

The four sequences marked in red initially scored as recKoRV1 because of the presence of PhER and KoRV in the sequences were not confirmed as recKoRV1 by Sanger sequencing but rather recKoRVs with different recombination sites.